When Index Term Probability Violates the Classical Probability Axioms Quantum Probability can be a Necessary Theory for Information Retrieval

Massimo Melucci
University of Padua
Italy

m.melucci@acm.org

Abstract

Probabilistic models require the notion of event space for defining a probability measure. An event space has a probability measure which ensues the Kolmogorov axioms. However, the probabilities observed from distinct sources, such as that of relevance of documents, may not admit a single event space thus causing some issues. In this article, some results are introduced for ensuring whether the observed probabilities of relevance of documents admit a single event space. Moreover, an alternative framework of probability is introduced, thus challenging the use of classical probability for ranking documents. Some reflections on the convenience of extending the classical probabilistic retrieval toward a more general framework which encompasses the issues are made.

1 Introduction

In Information Retrieval (IR), probabilistic models are employed for estimating the probability that a relevance or generation relationship exists between the information conveyed by a document and a user's information need represented by a query or any other user's action, such as browsing, document retention or click-through. These models requires an event space, which consists of a set of events and a probability measure, thus two event spaces differ due to either the events or the probability distribution.

Sets of events are employed for representing the occurrence of documents, terms, queries, relevance, aboutness, and their inter-relationships; for example, the intersection of the set which represents a document and the set of relevance represents the event that the document is relevant. For every single event space, a probability distribution P exists such that, for any pair of events, we can write the conditional probabilities as

$$P(A|B) = \frac{P(A \cap B)}{P(B)} \tag{1}$$

The latter is known as Bayes' Postulate (BP), which is different from Bayes' Theorem (BT) which states that, for any pairs of events,

$$P(B)P(A|B) = P(B|A)P(A) \tag{2}$$

While BP is a postulate, BT permits to compute probability distributions through the probability update mechanism provided by (2). Note that BP implies BT but the *vice versa* does not hold. Finally, and mostly important, the events whose probabilities used to compute the conditional probabililities through BP belong to a single probability space.

Some simple numerical examples in [7, pages 320, 321] point out that there might be something wrong or imprecise when defining the event space. In particular, some contradictions were noted when conditional probabilities are estimated from distinct event spaces and are then combined together by using BP since as the events whose probabilities used to compute the conditional probabililities through BP come from a single probability space. It was noticed that BP cannot be applied without asking some questions and that some measurements expressed in terms of events or random variables are simply ill-defined.

The contradictions found in [7] can be briefly explained as follows. Suppose that we are given two conditional probabilities $P(A|B), P(A|C)$ calculated by using BP, and we are also told that A, B, C do not necessarily come from a single event space. In other words, these probabiilities are *estimated* under different *contexts* – $P(A|B)$ has been estimated under B and $P(A|C)$ has been estimated under a *different* context C where the conditioning events (i.e., B and C are the contexts). Hence a single measure P may not exist such that (1) holds [1].

Another question is whether BT may still hold when the probabilities are estimated under different contexts. Suppose one is provided with an estimation of $P_1(B|A), P_4(B|\underline{A}), P_0(A)$ from three different event spaces[1]. By summing the right-hand side of (2) over A, we have that

$$P_2(B) = P_1(B|A)P_0(A) + P_4(B|\underline{A})P_0(\underline{A}) \tag{3}$$

whose sum over B yields 1, the latter being called Law of Total Probability (LTP). Hence, the following probability distributions are provided:

- $P_0(A)$ from event space 0,
- $P_1(B|A)$ from event space 1,
- $P_2(B)$ from event space 2,
- $P_3(A|B)$ from event space 3,
- $P_4(B|\underline{A})$ from event space 4.

However, one cannot choose the probability distributions at any degree of freedom. Some inequalities constrain the set of admittable probability distributions. Consider $P_3(A|B) + P_3(\underline{A}|B) = 1$. Hence, for every $P_0(A)$, we have that $P_1(B|A) \leq P_2(B) \leq P_4(B|\underline{A})$ or $P_4(B|\underline{A}) \leq P_2(B) \leq P_1(B|A)$ or equivalently

$$\left\| \frac{P_2(B) - P_1(B|A)}{P_4(B|\underline{A}) - P_1(B|A)} \right\| \leq 1 \tag{4}$$

Inequality (4) is named statistical invariant in [1] and is necessary and sufficient condition for the existence of a valid $P(B)$. In particular, when the event space is supposed to be Boolean and then event intersection such as $A \cap B$ can be appropriately defined and observed, the violation of (4) tests the existence of a single event space equipped with event intersection. If (4) was not admitted, one cannot state that the observed $P_2(B), P_1(B|A), P_4(B|\underline{A}$ come from a single event space, and LTP is violated.

It seems that a violation of (4) is detriment to IR effectiveness because, one may suppose, a "wrong" model of the world cannot provide nothing but "wrong" results. In contrast to intuition, as it often happens, some experiments have shown that the violation can lead to improvements of retrieval effectiveness [5]. These results are briefly described in the following section.

[1] The subscripts suggest that the P's refers to different event spaces.

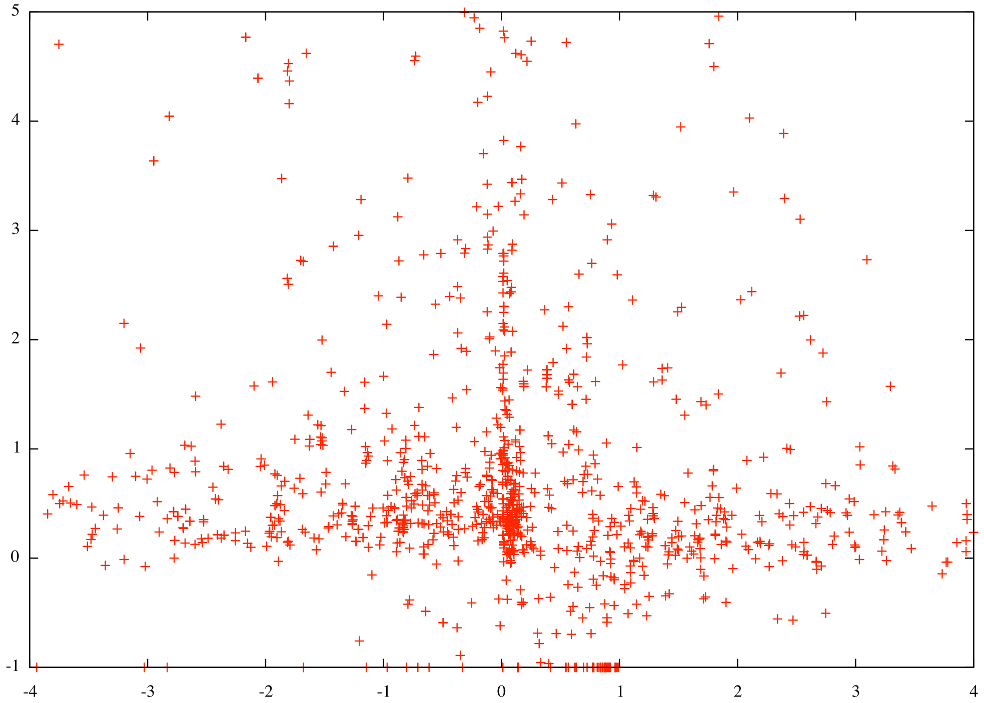

Figure 1: A scatterplot of (4) and the increase of MAP. Details in [5]

2 An Experiment Violating Some Probability Axioms

It was observed that the terms that correspond to B, C and violates (4), are those that increase average precision more frequently and significantly than those do not. (Event A means relevance.) The experiments reported in [5] aimed at answering the following question: "If the term suggested by the system to the user to expand the original query was so that the probability of occurrence violates the LTP, is the retrieval effectiveness measured on the new list of retrieved documents higher than that measured on the original list of retrieved documents?".

The experiments performed in [5] have suggested a relationship between the variation in Mean Average Precision and the statistical invariant 4 as depicted in Figure 1. A point of this plot corresponds to a value of fraction of (4) and an increase of MAP. The plot suggests that if the statistical invariant cannot be admitted and the LTP is violated, that is, if the selected term cannot be drawn from a sample whose probabilities have been estimated by aa single event space built from a set of relevant and non-relevant documents, then the increase in retrieval effectiveness may be observed and not only when the probability of occurrence of the selected term admits the invariant.

This experimental observation naturally lead to the question whether behind the violations of seemingly necessary theoretical invariants, are some potential for even further improving IR effectiveness. Something like this hypothesis has been formulated, for example, in [8] where the unification of logic, probability and vector space geometric within a single non-classical framework shapted by Quantum Mechanics (QM) has been suggested as a useful direction to this end. It follows that, other inequalities can be formulated that their violation can reveal, from the one hand, the inconsistency between experimental observations and the model, but, on the other hand, the directions toward a further development of probabilistic models for IR. The inequalities used throughout this article are illustrated in Appendix C and are based on [3]. In spite of the title, the contribution of [3] is not only an explanation of the role played by complex numbers in Quantum Mechanics, but it gives sufficient and necessary conditions for the admissibility of the conditional probabilities by a Classical Probability (CP) or Quantum Probability (QP) space. The topic was further investigated in [2].

3 Inequalities of Probability and IR

An IR system is trained on how to rank documents on the basis of the past interactions between the user and the system. These training data are recorded, for example, in a log-file or observed during a session in which the user searches for documents which fill his own information need. A notable example of training is Relevance Feedback and its several variations. When a system is trained by feedback the training data can be described as a table. Each tuple is an elementary event and one value is observed for each attribute and for each elementary event; for example, a tuple refers to a relevant document and the other attributes refer to the presence or absence of index terms[2].

[2] Tables is the standard way for representing training data as argued by I.H. Witten and E. Frank, *Data Mining: Practical Machine Learning Tools and Techniques*, Morgan Kaufmann, 2005.

1	2	3	4	5	6	7	8	9	10
A	$\underline{A}$	$\underline{A}$	$\underline{A}$	$\underline{A}$	A	$\underline{A}$	A	A	A
B	B	B	B	$\underline{B}$	B	$\underline{B}$	$\underline{B}$	$\underline{B}$	$\underline{B}$
C	C	C	C	C	$\underline{C}$	$\underline{C}$	$\underline{C}$	$\underline{C}$	$\underline{C}$

(a) Elementary events

A	A	A	A	$\underline{A}$	$\underline{A}$	$\underline{A}$	$\underline{A}$
B	B	$\underline{B}$	$\underline{B}$	B	B	$\underline{B}$	$\underline{B}$
C	$\underline{C}$	C	$\underline{C}$	C	$\underline{C}$	C	$\underline{C}$
1	1	0	3	3	0	1	1

(b) Event space

Figure 2: Event space for three observables.

For the sake of clarity and for making the illustration close to an IR scenario, three observables A, B, C are considered in this section; one can imagine that A refers to a set of relevant documents, B refers to the set of documents indexed by a term and C refers to the set of documents indexed by another term. This simplification does not eliminate the generality of the results because it will in the following be shown that some important properties will be violated even in this simple scenario, and therefore will in general be violated. At any rate, the most general case will be addressed in Section 4. The set of tuples of the table which stores the training data can represent the event space. Therefore, A is both an attribute and the subset of elementary events with that attribute; similarly, $A \cap B$ is another subset of elementary events.

An example is illustrated in Figure 2. The set of tuples of Figure 2(a) for which both terms occur, i.e., $\{1, 2, 3, 4\}$, is an event labeled A, B, C and placed in the first column of Figure 2(b); the last number of a column of Figure 2(b) is the cardinality of this event and it is used for computing the probability that both terms occur – the tuples that do not occur (e.g. $A, \underline{B}, C$) have null probability. When moving to the conditional probabilities, $p = P(B|A)$ is the probability that a term has been observed in a relevant document, $r = P(C|A)$ is the probability that the other term has been observed in a relevant document, and $q = P(C|B)$ is the probability that a term has been observed when the other term has also been observed. Suppose also that the frequencies of Figure 2(b) are used for estimating the probabilities by using BP. The example of Figure 2 obviously is a single event space and indeed Inequality 11 holds.

Inequalities might be violated if estimation is based on different event spaces. In IR, the violation of Inequality 11 may happen, for example, when

1. p, q, r are estimated using a mixture
2. p, q, r are estimated from distinct collections, or
3. the value of an observable is missing in the data for a few tuples.

When the probabilities are estimated by a mixture of, say, frequencies and additional knowledge, different sources are combined for estimating the probabilities p, q, r. This may happen, for example, in the Language Modelling (LM) approach when other sources of evidence, such as additional log-files or collection term frequency distributions, are exploited for adding some parameters and for smoothing probabilities. A linear combination is a well-known method for smoothing probabilities thus obtaining p, q, r as follows:

$$p = \alpha\frac{1}{2} + (1-\alpha)\frac{3}{4} \qquad q = \beta\frac{1}{2} + (1-\beta)\frac{1}{4} \qquad r = \gamma\frac{1}{2} + (1-\gamma)\frac{9}{15}\ . \tag{5}$$

Suppose that

$$\alpha = \frac{1}{9} \qquad \beta = \frac{1}{9} \qquad \gamma = \frac{2}{17}$$

By using Inequality 11, one can check that p, q, r do not admit a single event space and BT cannot be applied for computing $P(A|B)$ and $P(A|C)$.

A similar case happens when p, q, r are estimated from distinct collections; in this case, the experimental conditions yielding the probabilities are different and cannot be compared although the relevance assessments were given in the best possible way. This may happen, for example, when a document is stored in a collection, another document is stored in another collection and the query is routed to both these collections for retrieving and ranking the two documents in a single list. One may also think about routing the query to a collection drawn at random, or about merging the results received from a collection with those received from another after weighing the probabilities of relevance used for ranking the single lists.

Suppose, for example, that ten documents are stored in collection S_1 and other ten documents are stored in collection S_2 — Figure 3 reports an

S_1										S_2									
1	2	3	4	5	6	7	8	9	10	1	2	3	4	5	6	7	8	9	10
A	A	A	A	A	$\underline{\mathrm{A}}$	$\underline{\mathrm{A}}$	$\underline{\mathrm{A}}$	$\underline{\mathrm{A}}$	$\underline{\mathrm{A}}$	A	A	$\underline{\mathrm{A}}$	A	A	$\underline{\mathrm{A}}$	$\underline{\mathrm{A}}$	$\underline{\mathrm{A}}$	$\underline{\mathrm{A}}$	A
B	B	$\underline{\mathrm{B}}$	$\underline{\mathrm{B}}$	$\underline{\mathrm{B}}$	B	B	B	$\underline{\mathrm{B}}$	$\underline{\mathrm{B}}$	B	B	$\underline{\mathrm{B}}$	$\underline{\mathrm{B}}$	$\underline{\mathrm{B}}$	B	B	B	$\underline{\mathrm{B}}$	$\underline{\mathrm{B}}$
C	$\underline{\mathrm{C}}$	C	$\underline{\mathrm{C}}$	$\underline{\mathrm{C}}$	$\underline{\mathrm{C}}$	$\underline{\mathrm{C}}$	$\underline{\mathrm{C}}$	C	$\underline{\mathrm{C}}$	C	$\underline{\mathrm{C}}$	C	$\underline{\mathrm{C}}$	$\underline{\mathrm{C}}$	$\underline{\mathrm{C}}$	$\underline{\mathrm{C}}$	$\underline{\mathrm{C}}$	C	$\underline{\mathrm{C}}$

Figure 3: Ten elementary events for which two observables were measured for each urn.

example. Let p_i, q_i, r_i be the three conditional probabilities observed from S_i for $i = 1, 2$. One can easily check that

$$p_1 = P(B|A) = \frac{2}{5} \qquad q_1 = P(B|C) = \frac{1}{5} \qquad r_1 = P(C|A) = \frac{2}{5}$$

and that

$$p_2 = P(B|A) = \frac{2}{5} \qquad q_2 = P(B|C) = \frac{1}{5} \qquad r_2 = P(C|A) = \frac{1}{5}$$

The three conditional probabilities estimated from S_i do admit a single event space for each i — indeed, one single measure $\mu(X)$ can be defined as the frequency of the elementary events in X for each subset X of S_i.

Suppose that a query has to be submitted to a broker which has to decide the S_i to which the query has to be routed. The broker may either pick a collection at random with probability α and then route the query, or to route the query to both of them and then weigh the probabilities with weight α. In both cases, the conditional probabilities observed from the two collections may be

$$p = \alpha p_1 + (1 - \alpha)p_2 \qquad q = \alpha q_1 + (1 - \alpha)q_2 \qquad r = \alpha r_1 + (1 - \alpha)r_2 \ .$$

In this way, p is an estimation of the probability that B is observed in a relevant document stored in the S_i's where α is the prior probability that S_1 is selected. Similarly, q, r are estimated as a mixture of the conditional probabilities estimated from the single collections.

By using Inequality 11, one can check that when $\alpha = \frac{1}{2}$, the conditional probabilities

$$p = \frac{4}{10} \qquad q = \frac{2}{10} \qquad r = \frac{3}{10}$$

$$\begin{array}{cccccccccccc}
A & \underline{A} & \underline{A} & \underline{A} & \underline{A} & A & \underline{A} & A & A & A & ? & ? \\
B & B & B & B & \underline{B} & B & \underline{B} & \underline{B} & \underline{B} & \underline{B} & B & \underline{B} \\
C & C & C & C & C & \underline{C} & \underline{C} & \underline{C} & \underline{C} & \underline{C} & C & \underline{C}
\end{array}$$

Figure 4: Twelve elementary events are another set of examples of the relevance ($\{A, \underline{A}\}$), the presence of a term ($\{B, \underline{B}\}$) and the presence of another term ($\{C, \underline{C}\}$) observed in a log-file. Two elementary events include missing or known values for relevance.

do not admit a single event space. As a consequence, one cannot compute, say, $P(A|B)$ from the probabilities of the broker by using BT even though BT could be used for the probabilities estimated from single S_i. The reason was that the probabilities p_i, q_i, r_i come out from distinct spaces which describe two different experimental conditions. However, two distinct S_i's may still permit the probability of relevance to be computed by mixing the p_i, q_i, r_i's if appropriate values of α are fixed.

Another situation when the conditional probabilities do not admit a single event space occurs in the event of unknown or missing values, as exemplified in Figure 4. Depending on how the conditional probabilities are estimated, a single event space holds or does not. An estimation may only involve ten elementary events for which either A or $\underline{A}$ is known so that

$$p = \frac{\mu(B \cap A)}{\mu(B)} = \frac{2}{5} \qquad q = \frac{\mu(B \cap C)}{\mu(C)} = \frac{4}{5} \qquad r = \frac{\mu(C \cap A)}{\mu(C)} = \frac{1}{5}$$

Another estimation may also involve the tuples for which neither A or $\underline{A}$ is known so that

$$p = \frac{2/10}{6/12} = \frac{2}{5} \qquad q = \frac{5/12}{6/12} = \frac{5}{6} \qquad r = \frac{1/12}{6/12} = \frac{1}{6}$$

When the latter estimation is used, Inequality 11 is violated and therefore the single event space cannot hold. This outcome is little surprising when using the table of Figure 4 because the universe of the elementary events has two tuples for which a value is missing or unknown. A shrewd experimenter will avoid such a situation, yet it should be noted that a great deal of attention should be paid when conditional probabilities are provided by some "black-box" device or when the dataset includes missing values.

In the rest of this section, an example more extensive than the three-observable toy example is reported. The example aims at illustrating how the mathematical concepts described above can arise when designing an information retrieval experiment. In particular, it is shown how the estimation of a prior probability is a crucial step. The use of a small collection, such as the CACM of this example, is not detrimental to the generality of the results because even a small test collection or an experiment setting may contain a counter-example which invalidates the hypothesis of single event. The use of a larger collection would have provided more counter-examples.

The test collection was indexed so as to relate each one-keyword term to the documents in which it occurs. Before computing term frequencies, the stopwords provided with the test collection were removed from the documents. No stemming was computed. After indexing, the probability of relevance was first computed for each query as the relative frequency between the number of relevant documents and the total number of documents. Second, the probability of observing a term was computed as the relative frequency between the number of documents indexed by the term and the total number of documents. Third, the terms whose probability of occurrence was equal to the probability of relevance were selected, for each query. Finally, the probability of co-occurrence of every pair of two selected terms was computed, for each query. For example, the term `infinity` occurs in four documents and co-occurs with `translates` in one document; both terms have the same probability of occurrence as the probability of relevance to queries `23` or `30` which have four relevant documents.

In terms of probability, the universe of elementary events was the collection of documents. For each document, the property "the document was indexed by a term" and the property "the document is relevant to a given query" were observed; specifically, the latter property or observable was A, whereas B, C refer to two terms observed in a document. Therefore, the probability of relevance was $P(A)$ and the probability of observing a term was $P(X)$ where $X = B$ or $X = C$. From these probabilities, the conditional probabilities were computed. Specifically, the probability that a relevant document was indexed by C, i.e. $P(C|A)$ was computed as the relative frequency of relevant documents indexed by C. In the same way, the conditional probability of co-occurrence, i.e. $P(B|C)$ was computed as the relative frequency of documents indexed by C were indexed also by B. For example, only one relevant document (#`2786`) was indexed by `infinity` and therefore $P(C|A) = \frac{1}{4}$.

A sample of the results is reported in Table 1. The first row of the table includes a simple example which admits a single event space since, whenever

Query	Terms		$P(B\|C)$	$P(B\|A)$	$P(C\|A)$	CS	ReQS	CoQS
33	nonnormal	attainable	1	1	1	Yes	Yes	Yes
30	infinity	typesetting	$\frac{1}{4}$	$\frac{1}{4}$	$\frac{1}{2}$	Yes	No	Yes
30	translates	infinity	$\frac{1}{4}$	$\frac{1}{4}$	$\frac{1}{4}$	No	Yes	Yes
37	registers	compatible	$\frac{1}{12}$	$\frac{1}{12}$	$\frac{1}{12}$	No	No	No

Table 1: A sample of the conditional probabilities computed for the terms selected from the CACM test collections is reported in this table.

a term is observed in a document, it is certain that the other term or relevance is observed too. Given a query, two terms whose probability of occurrence equals the probability of relevance are associated with the conditional probability of co-occurrence and the conditional probability of relevance. Classical Space (CS) stands for "it admits a Classical Space", ReQuantum Space (QS) stands for "it admits a real QS" and CoQS stands for "it admits a complex QS". When a term admits a single event space, it admits a complex QS too, yet a real one may not be admitted as shown by the second row[3]. The probabilities of the third row admit a real QS, and then a complex one, while the single event space is not admitted. The last one is an instance that a QS can not be sufficient and that another space should be defined.

Let us concentrate on the third row. If one compiled a table which enumerates all the elementary events — one event for each document — then each row of that table would include a binary value which indicates the presence/absence of a term, another binary value which indicates the presence/absence of the other term, the document identifier and a binary value which indicates whether the document is relevant or not. If such a row is added for each document, an N-row table is obtained, where N is the total number of documents. One would easily check that a probability measure can be defined and therefore would find that a single event space can be defined. So, why does Table 1 say that K is not admitted? The reason is due to the fact that Accardi's inequalities described in Appendix C can be applied only if $P(A) = P(\underline{A}) = \frac{1}{2}$ (and similarly states the same for B and C), the latter being a hypothesis valid only if $N = 8$. This means that those

[3] QP spaces are introduced in Appendix B and are here mentioned to stress the fact that three observed conditional probabilities may admit a variety of theories of probability and not only the single event.

inequalities must be checked only if $P(A) = P(\underline{\mathrm{A}}) = \frac{1}{2}$.

4 Inequalities of Probability and the Probability Ranking Principle

What is the impact of the results illustrated in the previous sections on the PRP? Suppose that two documents have to be ranked. One document is represented by B, the other by C. When the observed conditional probabilities do not admit a single event space, the probability of relevance of B cannot be confronted to the probability of relevance of C. When these two probabilities cannot be confronted, the ranking is questionable. An explanation is that the probabilities are hinged on different measures, each defined by a distinct single event space which is an abstraction of a distinct experimental condition. In order to make the ranking sound, a single measure is needed, which is provided by a single probability space such as a QS. However, a single probability space is not sufficient for applying BT — for example, BP cannot be invoked if a QS is used. These issues are addressed in the rest of the article; in particular, the general case of an arbitrary number of documents to be ranked and then the impact on the PRP are investigated.

Let us now consider the general case. Suppose $m = n - 1$ documents are to be ranked by the probability of relevance — relevance is then the n-th observable. The events corresponding to the documents are $D_1, \ldots, D_m$, whereas the event corresponding to relevance is A, so the estimated probabilities are $P(D_i|A)$ for all $i = 1, \ldots, m$. The correlation vector of probabilities $\mathbf{p}$ used in Pitowsky's Theorem D.1 can be built as follows:

$$\begin{array}{rcll} p_i & = & P(D_i) & \\ p_{i,n} & = & P(D_i|A)P(A) & \qquad i = 1, \ldots m \\ p_n & = & P(A) & \end{array}$$

Now, suppose that $\mathbf{p}$ does not admit a single event space, that is, there is no set of values of $\lambda_1, \ldots, \lambda_{2^n}$ such that $\mathbf{p}$ can be expressed as a linear combination of fixed correlation vectors built from the 2^n binary strings (see Appendix D). Suppose, then, a subset of the n probabilities (or events) is selected and that the $n-1$ probabilities $p(1), \ldots, p(m-1), p(n)$ are considered. As a consequence, the bivariate probabilities $p(i,j)$ are computed from the corresponding events. In total, a $(n-1)n/2$-dimensional correlation vector

$\mathbf{p}'$ is observed. If this correlation vector admits a single event space, then two single event spaces will be found since the probability left apart admits one distinct space. If $\mathbf{p}'$ does not admit it, then the other similar correlation vectors are built by leaving one of $p(1), \ldots, p(m-1)$ apart at a time. If no $(n-1)n/2$-dimensional correlation vector admits a single event space, then the process is repeated by leaving two events apart until a single event space is admitted for $n-2$ events. The probability of relevance $p(n) = P(A)$ can be left last. It is known that the process will certainly end finding a single event space because the single $p(i)$ does always admit it. This means that eventually every document or relevance may be represented as an observable of a space being distinct from all the spaces which represent the other documents or relevance.

Corollary 4.1 *Let $\boldsymbol{p}$ be a correlation vector for (not necessarily disjoint) n events which does not admit a single event space. Then, there exist at least two subsets of events whose probabilities admit a single event space.*

5 Conclusions and Future Developments

The main conclusion which can be drawn from Corollary 4.1 is that, whenever one has to rank m documents by probability of relevance, a great deal of attention should be paid as to whether these probabilities admit a single event space or not. As mentioned above, the ranking of documents whose probability of relevance is computed from distinct spaces should be treated with due caution because the presence of distinct spaces would signal the use of different experimental conditions in which the probabilities of relevance were computed. Even if one decided to compare these probabilities despite the presence of distinct spaces, some properties of single event spaces, such as BT or distributivity, cannot be used since they are grounded on a single space.

It is our opinion that a new canvas is needed for probabilistic retrieval which overcomes the problems arising when some observed conditional probabilities cannot admit a single event space which is at the basis of the classical probabilistic retrieval models. Such a canvas should be based on QP for some reasons explained in the rest of this section.

In IR, it was conjectured that “if IR models are to be developed [...], without further empirical evidence to the contrary it has to be assumed that subspace logic will be non-classical.” [8]. The conjecture is the same as that made in Quantum Mechanics where Hilbert's spaces are taken as the theoretical framework for explaining how the phenomena studied by Physics happen in Nature at the particle-level. Whenever two observables A, B are measured on a particle, the event $A \cap B$, that is, the event that both A and B occurs, often does not make sense, that is, the design of an experiment which can determine A and B cannot in principle be implemented because the measurement of A interpheres with the measurement of B and thus nothing can be said about their co-occurrence.

If non-classical logic has to be assumed in IR too, the intersection of events cannot be assumed and as a consequence the events observed in IR cannot derive from the co-occurrence of events and then modeled as the intersection of sets. For example, one cannot observe both relevance and document as an event like $A \cap B$.

Nonetheless, the phenomena usually dealt with in Physics at the particle-level do not at first sight occur in IR and the proposal of using Hilbert's spaces does not imply that IR systems exhibit a quantum behaviour; it rather means that the mathematical framework is general enough for describing documents, queries and the retrieval of relevant information in a comprehensive way.

For measuring the uncertainty of the observation when non-classical logic has been supposed, QP has been suggested. A probabilistic retrieval function based on QP was illustrated in [4] where a new model for information retrieval was proposed for capturing the contextual properties being hidden in an object managed by an IR system. According to that proposal, the contextual properties are modeled as bases of a complex vector space and each value, called contextual factor, taken by these properties is modeled as one of the basis vectors. The probability that a contextual factor occurs in an object was modeled as the square of the inner product between the vector which represents the factor and the vector which represents the object, and was termed as probability of context. However, some questions were left unanswered.

An unanswered question in [4] was why the abstract vector spaces would be a better framework than other mathematical theories. A possible answer was provided in [8]: Hilbert's spaces encompass different models for information retrieval, such as the probabilistic model and the VSM. That answer stemmed from the conjecture that the use of non-classical logic is reasonable unless there is some evidence to the contrary. That notwithstanding, it was our opinion that it is still unclear why such a conjecture is necessary, and as a consequence, why QP is necessary in IR and a further explanation was necessary.

The results of the previous sections may provide an explanation of why QP is necessary in IR, that is, some conditional probabilities do not admit a single event space and therefore Bayes' results cannot be applied. Although QP is necessary, it might unfortunately not be sufficient.

References

[1] L. Accardi. On the probabilistic roots of the quantum mechanical paradoxes. In S. Diner and L. de Broglie, editors, *The wave-particle dualism*, pages 297–330. D. Reidel pub. co., 1984.

[2] L. Accardi. *Urne e camaleonti*. Il Saggiatore, 1997. In Italian.

[3] L. Accardi and A. Fedullo. On the statistical meaning of complex numbers in quantum mechanics. *Lettere al nuovo cimento*, 34(7):161–172, June 1982.

[4] M. Melucci. A basis for information retrieval in context. *ACM Transactions on Information Systems*, 26(3), 2008.

[5] M. Melucci. An investigation of quantum interference in information retrieval. In *Proceedings of the Information Retrieval Facility Conference (IRFC)*, 2010.

[6] I. Pitowsky. *Quantum Probability – Quantum Logic*. Springer-Verlag, 1989.

[7] S.E. Robertson. On event spaces and probabilistic models in information retrieval. *Journal of Information Retrieval*, 8(2):319–329, 2005.

[8] Keith van Rijsbergen. *The geometry of information retrieval*. Cambridge University Press, UK, 2004.

A Classical Probability

According to the CP, the events observed during an experiment[4] or in the real world are modeled as *sets*. A set is nothing other than an abstraction of an event, that is, every event corresponds to a subset of a larger, perhaps infinite, universe of elementary events. In this framework, the theory of sets together with its operations are an abstraction of the different ways the events can occur in the real world. In particular, the intersection of two sets models the conjunction, that is, the co-occurrence of two events, while the complement of a set models the negation of an event.

When probability is entered, the sets used for modeling the events of a real world or an experiment are subjected to a measure, that is, a real function of the sets. The measure is then used for computing the probability of the events modeled by the sets. In this way, the uncertainty of the occurrence of an event can effectively be measured. The event space together with the probability give rise to a probability space and in particular single event space. A notable example of the measure of a set (event) is the frequency of elementary events included by the the set (event), while its probability is the relative frequency.

Let us consider a mathematical formulation of the probability space, which will make things easier to illustrate in the remainder of this article. According to CP, the events $A, B, \ldots$ are subsets of the event space Ω and therefore $A \cap B$, $A \cup B$ and $\underline{\mathrm{A}} = \Omega \setminus A$ are subsets of the event space, too. Moreover, suppose $\mu : 2^{\Omega} \to [0, 1]$ is a measure of these subsets such that

$$\mu(\emptyset) = 0 \qquad 0 \leq \mu(A) \leq 1 \qquad \mu(\Omega) = 1 \ .$$

Note that μ is not necessarily a relative frequency. Moreover,

$$P(A) = \mu(A) \qquad P(A \cup B) = P(A) + P(B) \qquad \text{if} \qquad A \cap B = \emptyset$$

According to BP, which requires the single event space, the conditional probability of A given B is defined as:

$$P(A|B) = \frac{\mu(A \cap B)}{\mu(B)} \ . \tag{6}$$

[4] Experiment is meant in the broad sense of a procedure for gathering a set of data, or the set of data itself, in the context of testing a hypothesis or studying a phenomenon.

Bayes' Theorem (BT) states that

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}$$

The use of BT entails BP and then a single event space, and viceversa. In IR, the probabilities are often estimated through BT which combines the probability that, for example, a query is generated by the document and that the query is generated "a priori" by the collection or another resource.

B Quantum Probability

When Hilbert's spaces are considered, the events observed in an experiment or in the real world are modeled as *subspaces*, that is, every event corresponds to a subspace of a larger, perhaps infinite dimensional, Hilbert's space. As in CP, a subspace is nothing other than an abstraction of an event, and the theory of Hilbert's spaces together with its operations are an abstraction of the different ways the events can occur in the real world. According to that framework, documents, terms, queries, relevance, aboutness are events modeled as subspaces. While sets and CP are widely used in IR, subspaces are not employed at all for modeling events. An exception is [8] where Hilbert's spaces were introduced for establishing a new theoretical framework encompassing different models proposed by the IR community over the years.

What is important to note is that the operations commonly defined for sets are not always defined for subspaces; for example, the notion of union thought for sets cannot be defined on subspaces because the union of two subspaces is not a subspace, while the linear span of two subspaces is. The opposite holds too, i.e., the linear span which is defined for subspaces cannot be defined for sets. However, there may be analogies — the complement of a set corresponds to an orthogonal subspace and both may model the negation of an event, yet the complement of a subspace is not the same as the complement of a set: the former is the set of vectors orthogonal to the vectors of the set, the latter is the set of vectors not in the set.

When probability is entered, the subspaces used for modeling the events of a real world or an experiment are subjected to a measure, that is, a real function of the subspaces of a Hilbert's space. The measure is then used for computing the probability of the events modeled by the subspaces. In this way, the uncertainty of the occurrence of an event can be measured even when sets are replaced with subspaces. In the QP case, therefore, the probability space is given by subspaces and a "QP" function, as illustrated in the following.

As before, let us consider a mathematical formulation. According QP, the events $X, Y, \ldots$ are modeled as subspaces $|X\rangle, |Y\rangle, \ldots$ of the space.[5] The definition of the probability of an event modeled by a subspace is different from that of the same event but modeled by a subset. Indeed, the probability of an event modeled by a subspace depends on a subspace and can then be considered as a conditional probability defined as

$$P(X|Y) = |\langle Y|X\rangle|^2 \ . \tag{7}$$

where the modulus of the inner product between the two vectors is called *amplitude*. In IR terms, Equation 7 is the probability that, say, a document described by index term Y is relevant. The same applies when $X, Y, Z, ...$ models, for example, terms, aboutness, or document clusters.

The conditional probabilities can be defined for both CP and QP, yet they are defined in different ways. What is different is that QPs are inherently conditional because the probability of an event is conditioned to a subspace, which refers to another event. Moreover, the conditional probabilities in QP theory are symmetric since

$$P(X|Y) = |\langle Y|X\rangle|^2 = |\langle X|Y\rangle|^2 = P(Y|X) \ .$$

Therefore, the conditional probabilities used in the article become

$$p = |\langle A|B\rangle|^2 \qquad q = |\langle B|C\rangle|^2 \qquad r = |\langle A|C\rangle|^2 \tag{8}$$

C Inequalities of Probability

Suppose there are three observables with n possible values each. Without loss of generality, and for making illustration easier, it is assumed that $n = 2$, that is, each observable has two mutually exclusive values, e.g., $\underline{A}, A$. As for IR, B might be the event that a term is observed in a document, while $\underline{B}$ is

[5] What follows concentrates on one-dimensional subspaces, that is, on the set of vectors spanned by a vector $|X\rangle$ which models an event A — from now on, $|Y\rangle$ means the subspace too. $|.\rangle$ is called bra-ket notation and was introduced by P. Dirac in *The Principles of Quantum Mechanics*, Oxford University Press, 1958.

the event that the term is not observed and A ($\underline{A}$) may denote the event that a document is (not) relevant and so on. Of course, three observables only is a small example, but the fact that the single event space cannot be admitted even when only three observables are examined suggests that it cannot be admitted in more general cases. It is assumed that

$$\begin{aligned} p = P(A|B) = P(B|A) \\ q = P(B|C) = P(C|B) \\ r = P(A|C) = P(C|A) \end{aligned}$$

as in [3]. When CP is used, the symmetry of the conditional probability implies that

$$\mu(A) = \mu(B) = \mu(C) \ .$$

This "symmetry" may well happen in an IR context; for example, the probability that a given index term is chosed for a relevant document, that is, $P(C|A)$, may equal the probability that the document is assessed as relevant if it has been indexed by the index term. The fact that the conditional probabilities are usually asymmetric is due to the use of CP and then of BP which in fact makes $P(A|B)$ different from $P(B|A)$. When CP is used, one measure μ exists such that

$$p = \frac{\mu(A \cap B)}{\mu(B)} \qquad q = \frac{\mu(B \cap C)}{\mu(C)} \qquad p = \frac{\mu(A \cap C)}{\mu(C)} \ . \tag{9}$$

Suppose, for example, that $p = q = r = \frac{1}{2}$ and $\mu(A) = \mu(B) = \mu(C) = \frac{1}{2}$. It can easyly be seen that the measures of the co-occurring events can be computed as

$$\mu(A \cap B) = \mu(B \cap C) = \mu(C \cap A) = \frac{1}{4} \ .$$

Things change when other values of p, q, r are estimated from sources independent of each other, for example, as

$$p = \frac{13}{18} \qquad q = \frac{5}{18} \qquad r = \frac{10}{17} \tag{10}$$

The surprising result is that when the values of Equations 10 are considered, a measure μ cannot be defined in a way such that the probability of $A \cap B$,

$B \cap C$, $C \cap A$ and $A \cap B \cap C$ exist — no single event space can admit those values of p, q, r. The values of Equations 10 are not the only possible values and an infinite number of values of p, q, r exist such that the events do not admit a meausure according to CP. As CP estimates the conditional probabilities on the basis of BP, it follows that,

$$P(A|B) \neq \frac{\mu(A \cap B)}{\mu(B)}$$

since $\mu(A \cap B)$ cannot be calculated. If the co-occurrence of events, e.g. $A \cap B$ was observed, and the frequency of these co-occurring events were available, an estimation of $\mu(A \cap B)$ would be possible thus making BP valid. When μ cannot be defined for some observed conditional probabilities, one has to conclude that the co-occurrence of events is impossible, namely, statements like "both A and B occur" do not make any sense. The inequality which acts as the test of the existence of a single event is proven in [3] and is stated as

Proposition C.1 *p, q, r admit a single event space if and only if*

$$|p + q - 1| \leq r \leq 1 - |p - q| \tag{11}$$

When Inequality 11 is violated, neither measure μ nor sets $A, B, C, ...$ can be defined for the observables A, B, C such that BP holds.

Inequality 11 provides for a simple test to check if the conditional probabilities estimated in different experimental conditions are compatible with a single event space. The question is then, what is the probability space if that single event is incompatible? The answer was provided in [3] and is reported here without proof.

Proposition C.2 *p, q, r admit a complex QP space if and only if*

$$\left(\sqrt{pq} - \sqrt{1-p}\sqrt{1-q}\right)^2 \leq r \leq \left(\sqrt{pq} + \sqrt{1-p}\sqrt{1-q}\right)^2 \tag{12}$$

Proposition C.3 *p, q, r admit a real QP space if and only if*

$$r = \left(\sqrt{pq} - \sqrt{1-p}\sqrt{1-q}\right)^2 \qquad \textit{or} \qquad r = \left(\sqrt{pq} + \sqrt{1-p}\sqrt{1-q}\right)^2 \tag{13}$$

From these inequalities, some simple results follow. First, if p, q, r admit a single event space, then they also admit a QS. Second, there are infinite values of p, q, r which admit a complex QS, and not a single event space. When p, q, r admit a complex (real) QP space, the events A, B, C can be represented as vectors $|A\rangle, |B\rangle, |C\rangle$ of a complex (real) QS such that the probabilities are those defined in Section B. There are cases when either the complex QS nor the CP space can be admitted; for example, $p = 1/10, q = 2/10, r = 3/10$. This implies that that QS is a necessary yet not sufficient framework.

Up to now, the relatively simple case of three observables or properties has been considered. A more general result states a necessary and sufficient condition that a set of probabilities admit a single event space. This result is due to Pitowsky, was proven in [6] and is illustrated in Appendix D.

D Pitowsky's Theorem

Suppose $n \geq 2$ properties are observed from a, say, collection of documents — in particular, the case $n = 3$ was considered in the previous sections where the properties were labeled as A, B, C and their respective negations $\underline{\mathrm{A}}, \underline{\mathrm{B}}, \underline{\mathrm{C}}$ to mean, for example, that a document was relevant (A) or not ($\underline{\mathrm{A}}$).

Suppose also that a series of experiments yielded the $n(n+1)/2$ probabilities

$$p(1), \ldots, p(n), p(1,1), \ldots, p(i,j), \ldots, p(n-1,n)$$

where $1 \leq i < j \leq n$, where p_i is the probability that the event A_i occurs and $p_{i,j}$ is the probability that events A_i, A_j are observed in the series of experiments. These probabilities can be arranged in the correlation vector $\mathbf{p}$. Given $\mathbf{p}$, under what conditions a single single event space for the events $A_1, \ldots, A_n$ and the measure μ can be defined such that for all i, j where $1 \leq i < j \leq n$

$$p(i) = \mu(A_i) \qquad p(i,j) = \mu(A_i \cap A_j) \text{ ?}$$

The answer was provided in [6].

Suppose that n properties are considered and that all the strings of n binary numbers 0's and 1's are built. Let b be one of these binary strings, i.e. $b \in \{0,1\}^n$; for example, when $n = 2$, then $b = \mathtt{01}$ is such a string. The binary digit `1` means that A_i was observed, while `0` means that it was not.

Once b is fixed, the correlation vector of $n(n+1)/2$ probabilities $\mathbf{p}_b$ is defined as follows: $\mathbf{p}_b(i) = b_i$ and $\mathbf{p}_b(i,j) = b_i b_j$; for example, when $b = \texttt{01}$, then $\mathbf{p}_b = (0,1,0)$. There are 2^n such correlation vectors since 2^n binary strings can be enumerated using n binary digits.

The theorem defined the closed, convex polytope whose vertices are the 2^n correlation vectors like $\mathbf{p}_b$, that is, the polytope is the set of all points that can be expressed by a linear combination of these 2^n correlation vectors. As formula, the polytope is expressed as

$$\mathbf{p} = \lambda_1 \mathbf{p}_{b_1} + \cdots \lambda_{2^n} \mathbf{p}_{b_{2^n}}$$

where b_i is the i-th binary string, $\mathbf{p}_{b_i}$ is the correlation vector of probabilities built from b_i and

$$\lambda_i \geq 0 \qquad \sum_{i=1}^{2^n} \lambda_i = 1 \ .$$

Then, the following proposition holds

Theorem D.1 *For all n and all correlation vectors of probabilities $\boldsymbol{p}$, $\boldsymbol{p}$ admits a single event space of n (not necessarily distinct) events if and only if $\boldsymbol{p}$ belongs to the polytope.*

This means that the 2^n-unknowns system of $n(n+1)/2$ linear equations

$$\left\{ \begin{array}{ccccccc} p(1) & = & \lambda_1 p_{b_1}(1) & + & \cdots & + & \lambda_{2^n} p_{b_{2^n}}(1) \\ \vdots & = & & & \vdots & & \\ p(n) & = & \lambda_1 p_{b_1}(n) & + & \cdots & + & \lambda_{2^n} p_{b_{2^n}}(n) \\ p(1,n) & = & \lambda_1 p_{b_1}(1,n) & + & \cdots & + & \lambda_{2^n} p_{b_{2^n}}(1,n) \\ \vdots & = & & & \vdots & & \\ p(m,n) & = & \lambda_1 p_{b_1}(m,n) & + & \cdots & + & \lambda_{2^n} p_{b_{2^n}}(m,n) \end{array} \right. \tag{14}$$

have solutions if and only if $\mathbf{p}$ belongs to the polytope.

Note that since the events are not necessarily distinct, one can assume that when $n = 2$, $A_1 = A_2$ and therefore $p(1) = p(2) = p(1,2)$. In this way, the theorem always holds since $\mu(A_1) = p(1)$ and the event space includes only one observable.

The theorem holds if and only if a system of inequalities holds. For example, suppose $n = 2$, $n(n+1)/2 = 3$ and $2^2 = 4$. The binary strings are $b_1 = \texttt{000}, b_2 = \texttt{010}, b_3 = \texttt{100}, b_4 = \texttt{111}$. Three observed probabilities $p(1), p(2), p(1,2)$ admit a single event space if and only if the system:

$$\begin{cases} 0 \le p(1,2) \le p(1) \le 1 \\ 0 \le p(1,2) \le p(2) \le 1 \\ 0 \le p(1) + p(2) - p(1,2) \le 1 \end{cases}$$

admits solutions. For example, no solutions can be admitted when $p(1,2) > p(1)$ or $p(1,2) > p(2)$. When $n = 3$,

$$\begin{cases} 0 \le p(i,j) \le p(i) \le 1 & 1 \le i < j \le 3 \\ 0 \le p(i) + p(j) - p(i,j) \le 1 & 1 \le i < j \le 3 \\ p(1) - p(1,2) - p(1,3) + p(2,3) \ge 0 & \\ p(2) - p(1,2) - p(2,3) + p(1,3) \ge 0 & \\ p(3) - p(1,3) - p(2,3) + p(1,2) \ge 0 & \end{cases}$$

The main problem is that a polynomial-time algorithm which tests if $\mathbf{p}$ belongs to the polytope for every n does not exist. However, in this article, our interest is not to compute the polytope, but to have theoretical results which provide the necessary and sufficient conditions so that $\mathbf{p}$ admits a single event space.